Statistical Self-Similar Properties of Complex Networks

Chang-Yong Lee*

The Department of Industrial Information,

Kongju National University, Chungnam, 340-702 South Korea

Sunghwan Jung

Applied Mathematics Laboratory, Courant Institute for Mathematical Science,

New York University, New York, NY 10012 USA

Abstract

It has been shown that many complex networks shared distinctive features, which differ in many ways from the random and the regular networks. Although these features capture important characteristics of complex networks, their applicability depends on the type of networks. To unravel ubiquitous characteristics that complex networks may have in common, we adopt the clustering coefficient as the probability measure, and present a systematic analysis of various types of complex networks from the perspective of statistical self-similarity. We find that the probability distribution of the clustering coefficient is best characterized by the multifractal; moreover, the support of the measure had a fractal dimension. These two features enable us to describe complex networks in a unified way; at the same time, offer unforeseen possibilities to comprehend complex networks.

PACS numbers: 89.70.+c, 05.45.Df, 87.23.Ge

*Electronic address: clee@kongju.ac.kr

I. INTRODUCTION

The structure of complex systems across various disciplines can be abstracted and conceptualized as networks (or graphs) of nodes and links to which many quantitative methods can be applied so as to extract any characteristics embedded in the system [1]. Numerous complex systems have been expressed in terms of networks, and have often been categorized by the research field, such as social [2, 3], technological [4, 5], and biological networks [6, 7], to name a few.

It was shown that many complex networks had distinctive global features in common, including the small-world [8] and the scale-free [9] properties. These uncovered characteristics distinguish complex networks from the random and the regular networks in that the average path between any two nodes is shorter while maintaining highly clustered connections, and that the degree of nodes approximately follows a power law distribution. In addition to the global characteristics, investigations on complex networks at the local level have been directed to reveal local characteristics from the perspective of finding patterns of interconnections among nodes. Notable examples include the network motif [10], the (dis)assortativity [11, 12], and the topological modules in the metabolic [13] as well as the protein interaction networks [12]. Motivated by these characteristics, numerous models for the network growth and evolution have been proposed to understand the underlying mechanism of complex networks.

To gain further understanding of complex networks, we investigate local features of the networks from the perspective of the statistical self-similarity. This may provide us with not only deeper insight into complex networks, but a unified way of describing them. To this end, we focus on the clustering coefficient [8] of a node i, defined as

$$C_i = \frac{n_i}{k_i\ (k_i - 1)}\ , \tag{1}$$

where n_i, $0 \leq n_i \leq k_i(k_i-1)$, is the number of directed links (an undirected link counts twice) among k_i nearest neighbor nodes of the node i. It is a measure of the interconnectivity among nearest neighbors of a node, or the modularity [13], thus can be a quantity representing the local structure of the network.

The clustering coefficient has been analyzed from the perspective of the degree correlation. It was found that the clustering coefficient correlated with the degree in some complex

networks. In particular, there is a power law correlation between the clustering coefficient and the degree for the deterministic scale-free network model [14], the Internet [15], and metabolic networks [13]. It is a form of $\langle C \rangle(k) \sim k^{-\delta}$, where $\langle \cdot \rangle$ represents the average over the same degree, and δ depends on the type of networks. However, a similar analysis of the clustering coefficient reveals that the power law correlation is not manifest to other types of networks, such as the protein interaction and social networks. As shown in Fig. 1, the clustering coefficient of the film actor network correlates with the degree not with a power-law, but exponentially (Fig. 1A); while the neural network has an approximate linear correlation between the clustering coefficient and the degree (Fig. 1B). Moreover, there is no evident correlation in the protein networks (Fig. 1C, 1D).

This finding suggests that in general the clustering coefficient is not a simple quantity which can be related for some common ground to other quantities, such as the degree. Thus, it is desirable to analyze the clustering coefficient beyond the degree correlation. In this paper, we focus on the systematic analysis of the clustering coefficient in the wide classes of complex networks [16], together with theoretical models for the complex networks, such as the random [17], the scale-free [9], and the small-world networks [8].

II. DATA DESCRIPTION

We analyzed the clustering coefficient in the complex networks of the film actor network, WWW, the scientific collaboration network, metabolic networks, protein interaction networks, the neural network, and the Internet of both Autonomous System (AS) and the router levels; together with models for the random, the scale-free, and the small-world networks. For directed networks such as metabolic networks and WWW, we distinguish the network into the directionality (in-degree and out-degree) and carry out separate analysis. The source of the network data is in Ref. [16], and more information of each network is the following.

Film actor: An actor represents a node, and actors casted in the same movie are linked. (374511 nodes and 2445818 undirected links)

Scientific collaboration: Each author corresponds to a node and co-authors are linked among others in the Los Alamos E-print Archive between 1995 and 1999 (inclusive) in the field of the condensed matter. (13861 nodes and 89238 undirected links)

Internet of Autonomous Systems (AS) level: Each autonomous system represents a node, and a physical network connection between nodes represents a link. (6474 nodes and 25144 undirected links)

Internet of router level: The Internet connection at the router level. The data is collected by the Mercator (`http://www.isi.edu/scan/mercator/`), a program that infers the router-level connectivity of the Internet. (284772 nodes and 898456 undirected links)

WWW: World Wide Web connection network for `http://www.nd.edu`. Each HTML document represents a node, connected directionally by a hyperlink pointing from one document to another. It is a directed network of 325729 nodes and 1469679 directed links.

Metabolic networks: Metabolic networks of six organisms, two for each domain, are analyzed. They are *Archaeoglobus fulgidus* (459 nodes and 2155 directed links) and *Methanobacterium thermoautotrophicum* (399 nodes and 1937 directed links) in Archae; *Escherichia coli* (698 nodes and 3747 directed links) and *Salmonella typhi* (735 nodes and 3882 directed links) in Bacteria; *Saccharomyces cerevisiae* (511 nodes and 2690 directed links) and *Caenorhabditis elegans* (413 nodes and 2061 directed links) in Eukaryote. Note that the metabolic network is a directed network.

Protein interaction networks: We have analyzed protein interaction networks of six organisms. They are *Saccharomyces cerevisiae* (4687 nodes and 30312 undirected links), *Escherichia coli* (145 nodes and 388 undirected links), *Caenorhabditis elegans* (2386 nodes and 7650 undirected links), *Drosophila melanogaster* (6926 nodes and 41490 undirected links), *Helicobacter pylori* (686 nodes and 2702 undirected links), and *Homo sapiens* (563 nodes and 1740 undirected links)

Neural network: Somatic nervous system of *Nematode C. elegans* except that in the pharynz is considered. A link joins two nodes representing neurons, if they are connected by either a synapse or a gap junction. All links are treated as undirected. (265 nodes and 3664 undirected links)

III. MULTIFRACTALITY OF COMPLEX NETWORKS

The set of C_i for each network can be used to form a probability distribution (Fig. 2). As shown in Fig. 2A-2D, the distribution of C_i in complex networks differ considerably from that

of the random network (Fig. 2F). Probability distributions for complex networks bring out high irregularity of various intensities in different clustering coefficients, developing a long tail extending to either large (Fig. 2A-2C) or small (Fig. 2D) values of the clustering coefficient. This suggests that not a few but many, possibly infinite, parameters may be needed to characterize the distribution. To quantify the variation in the distribution, a continuous spectrum of scaling indices has been proposed [18]. For the spectrum that quantifies the inhomogeneity in the probability distribution, we utilize the clustering coefficient as the probability measure, and analyze the distribution from the perspective of the statistical self-similarity, the multifractal [19, 20].

The multifractal, which is not necessarily related to geometrical properties [21], is a way to describe different self-similar properties for different "regions" in an appropriate set (in our case, different values of the clustering coefficient), and applied, for instance, to the fully developed turbulence [21, 22], which is one of the most common examples of complex systems. It consists of spectra displaying the range of scaling indices and their density in the set, thus has been used to explain richer and more complex scaling behavior of a system than the case in the critical phenomena. The multifractal can be accomplished by examining the scaling properties of the measure characterized by the singularity strength α and its fractal dimension $f(\alpha)$, which roughly indicates how often a particular value of α occurs [18]. In practice, α and $f(\alpha)$ are customarily obtained from the Legendre transformation of q and D_q, via

$$\alpha = \frac{d}{dq}\{(q-1)D_q\} \ , \tag{2}$$

and

$$f(\alpha) = q\alpha - (q-1)D_q \ , \tag{3}$$

where D_q is the generalized correlation dimension often estimated by the correlation integral method [23]. It is the quantity for anomalous scaling law whose value depends on different moment q. D_q is defined as

$$D_q = \lim_{R\to 0} \frac{1}{q-1}\frac{\ln S_q(R)}{\ln R} \ , \tag{4}$$

where $S_q(R)$ is known as the correlation sum (or correlation integral), and it is given, using Heaviside function Θ, as

$$S_q(R) = \frac{1}{M}\sum_{j=1}^{M}\left\{\frac{1}{M-1}\sum_{k=1,k\neq j}^{M}\Theta\left(R - |C_j - C_k|\right)\right\}^{q-1} \ , \tag{5}$$

where M is the number of nodes and C_i is the clustering coefficient of node i. The spectrum D_q may be smoothed before transforming into α and $f(\alpha)$ to avoid the contradiction of the general property of D_q, i.e., $D_q \leq D_{q'}$ if $q' \leq q$.

There is a difficulty in taking the limit $R \to 0$ in Eq. (4) for a finite number of data points. Due to the finiteness, there always exists the minimum distance of $|C_j - C_k|$. Thus, when R is less than the minimum distance, the correlation sum becomes zero and no longer scales with R. Therefore, in practice, the generalized dimension D_q is determined by plotting $\ln S_q(R)/(q-1)$ as a function of $\ln R$ and estimating the slope within an appropriate scaling region using a least square fit. The error associated with the fit can be obtained as a statistical uncertainty based on fitting a straight line in the scaling region.

Fig. 3 displays the estimated $f(\alpha)$ versus α for various complex networks. As shown in Fig. 3A-3D, the infinitely many different fractal dimensions, manifested by the shapes of $f(\alpha)$, suggest that the measure is a multifractal, and thus, cannot be explained by a single fractal dimension. All of the complex networks we have examined, except for the neural network of *Caenorhabditis elegans*, form multifractals irrespective of their global characteristics, such as the number of nodes and their degree distributions. Furthermore, for all complex networks we have analyzed, the average and standard deviation of the most probable singularity strength α_0, where $f(\alpha)$ takes its maximum value, are $\langle \alpha_0 \rangle = 1.2 \pm 0.3$; those of $f(\alpha_0)$ are $\langle f(\alpha_0) \rangle = 0.8 \pm 0.1$.

The multifractal observed in complex networks implies that the distribution of clustering coefficients can be described as interwoven sets of singularities of strength α, each of which has its corresponding fractal dimension $f(\alpha)$ [18]. In our case, this implies that different values (or range of values) of the clustering coefficient may have different fractal dimensions. From the viewpoint of network dynamics in which rewiring and/or adding new nodes and links are involved, the multifractal suggests that as a network grows, nodes of large clustering coefficients change their clustering coefficients by a factor that differs from nodes of small clustering coefficients change theirs.

The different rate of changing the clustering coefficient may stem from two sources (or modes): the degree of a node k and the corresponding interconnectivity n. For a fixed k, the clustering coefficient depends only on the interconnectivity n, so that $C \sim n$. In this case, the dynamics (via rewiring and/or adding new links) drive networks in such a way that different values of n are not equally probable; rather, some values of n are more probable

than others. This assertion is further supported by the fact that as k increases, the number of distinct n does not increase quadratically in k, and that n and k are linearly correlated (see below). For a fixed n, $C \sim k^{-2}$. Thus, the addition of new links to higher degree nodes is more likely to drop their clustering coefficient much faster than that of new links to lower degree ones. Therefore, the dynamics of complex networks can be characterized by an evolution via interplay between the two sources.

Contrary to most complex networks, the irregularity of the distribution is absent in the neural network of *Caenorhabditis elegans* (Fig. 2E). The estimate $D_q \approx 0.83 < 1$ is independent of q, resulting in $\alpha = f(\alpha) \approx 0.83$. Thus, the measure is not a multifractal, rather it can be characterized by a single fractal dimension. The absence of the multifractality is probably due to the biologically intrinsic property of the neuron. The geometric character of the neuron imposes a constraint on the number of synaptic contacts (i.e. links), leaving no room for the irregularity of the distribution [2, 24].

Typically, $f(\alpha)$ satisfies $0 < f(\alpha) < D_0$, where D_0 is the dimension of the support of the measure, which is the set of clustering coefficients without their relative frequencies. We find $D_0 < 1$ for all complex networks (Fig. 3A-3D), indicating that supports of the measure have fractal dimensions, just like the Cantor set. This means that forbidden regions are embedded in the range of allowed clustering coefficients so that some values are highly suppressed. A possible cause of the suppression might stem from the correlation between the degree k of a node and its corresponding interconnectivity n. A simple statistic, such as the Pearson's correlation coefficient r, ranging $-1 < r < 1$, can be used to quantify the correlation, as it reflects to what extent the two quantities are linearly correlated. The result (Fig. 4A) reveals that complex networks have higher linear correlation than the random network. Moreover, some complex networks, such as metabolic networks and the Internet of AS level, disclose strong linear correlations ($r > 0.95$).

For metabolic networks and the Internet of AS level in which the degree of a node and its interconnectivity among its nearest neighbor nodes are strongly correlated, we ask how the next nearest neighbor nodes are interconnected, and whether the distribution of the corresponding clustering coefficients maintains the multifractality. To this end, we extend the definition of the clustering coefficient to the next nearest nodes by including the next nearest nodes for both the degree of a node k and its interconnectivity n, which is the number of links between two next nearest nodes. The irregularity and the long-tail characteristics

are again found in the distribution of the extended clustering coefficient, suggesting the existence of the multifractal. As shown in Fig. 4B-4D, the probability distribution of the extended version of the measure can again be characterized by the multifractal, indicating that the statistical self-similarity is not necessarily restricted to the local interconnectivity. By including the next nearest neighbor nodes to the definition of the clustering coefficient, more distinct values of the clustering coefficient are possible than that of the nearest neighbor nodes. This can be expected since the extended version of the clustering coefficient of a node includes an average over the clustering coefficients of its nearest neighbor nodes, partly smoothing out the irregularity. This is also manifested by the fact that D_0^*, the dimension of the support of the extended clustering coefficient, is bigger than corresponding D_0.

Based on the multifractal found in complex networks, for comparison, we carried out similar analysis to models of the random, the scale-free, and the small-world networks. In the case of the random network, the clustering coefficients are smoothly distributed, by having a "bell-shape" (Fig. 2F). Furthermore, $D_q \approx 1.0$ for all q, indicating that there is no self-similarity. This can be expected since the support of the measure can be regarded as a line, which is one dimensional.

The distribution of the clustering coefficient for the scale-free network shows the irregularity, similar to the case of complex networks; furthermore the distribution can be described by the multifractal (Fig. 3E). From simulation results with various different parameters for the network, however, we found that not only the most probable singularity strength α_0 ($0.64 < \alpha_0 < 0.76$), but the dimension of the support D_0 ($0.45 < D_0 < 0.50$) is smaller than that of complex networks, suggesting that more severe restriction is imposed on the possible values of the clustering coefficient. According to the model, nodes of higher degree are preferred to have additional links rather than those of lower degree. This preferential attachment leaves the clustering coefficient of high degree nodes to decrease much faster than that of low degree ones. Thus when the number of nodes is doubled, for instance, the clustering coefficient of high degree nodes changes by a factor different from that of low degree nodes, analogous to the kinetics of the diffusion-limited aggregation [25].

For the small-world network, the rewiring probability p dictates both the irregularity in the distribution and the multifractality. For a small rewiring probability (say, $p = 0.01$), the multifractal emerges (Fig. 3F); however the dimension of the support is $D_0 \approx 1.0$, implying that the set of the measure entirely covers the space of the clustering coefficient. As the

rewiring probability increases, the range of the clustering coefficients becomes smaller and the degree of inhomogeneity decreases, and then the network finally becomes a random network as we can easily anticipate.

IV. SUMMARY AND CONCLUSION

Based on the irregularity of intensities in the probability distribution of the clustering coefficient, we regarded the clustering coefficient as the probability measure and analyzed the clustering coefficient of various types of complex networks from the perspective of the statistical self-similarity. We found that the probability measure and the support of the measure can be characterized by the multifractal and the fractal, respectively. Furthermore, for complex networks having strong linear correlation between the degree and the interconnectivity, the multifractality extends into the clustering coefficient of the next nearest neighbor nodes. These characteristics are unique to the real complex networks and cannot be found in the random network. From the aspect of the multifractality, models of the scale-free and the small-world differ from real networks in the distribution of the singularity strength $f(\alpha)$.

The statistical self-similarity in the distribution of the clustering coefficient can be served as a general characteristic of complex networks; at the same time, giving a further insight into the understanding of complex networks. The multifractality shared by different complex networks suggests that a similar law may govern the diverse complex networks of nature as well as artificiality. Furthermore, it can be used to classify the complex networks, and serves as a "touchstone" of proposed models for complex networks.

Acknowledgments

We like to thank M. Newman for providing us with the scientific collaboration data. We also appreciate the open source of various complex network data available at `http://www.nd.edu/~networks/`. This work was supported by the Korea Research Foun-

dation Grant funded by the Korean Government (MOEHRD) (KRF-2005-041-H00052).

[1] For a review of the network theory, see, for example, M. Newman, SIAM Review **45** 167 (2003), and R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] L. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **97**, 11149 (2000).

[3] M. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).

[5] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

[6] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes, Nature Genetics **31**, 60 (2002).

[7] H. Jeong, B. Tombor, R. Albert, Z. Litvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).

[8] D. Watts and S. Strogatz, Nature (London) **393**, 440 (1998).

[9] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[10] R. Milo *et al*, Science **298**, 824 (2002).

[11] M.E.J. Newman, Phys. Rev. Letts. **89**, 208701 (2002).

[12] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[13] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002); E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).

[14] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **65**, 066122 (2002).

[15] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).

[16] The sources for various types of network data are the following. Film actors, WWW, and metabolic networks were obtained from `http://www.nd.edu/~networks/`; the scientific collaboration data was provided by M. Neuman; the Internet of Autonomous Systems level was obtained from `http://moat.nlanr.net/Routing/rawdata/`; the Internet of router level is collected by the Mercator and is available at `http://www.isi.edu/scan/mercator/`; protein interaction networks data are available at `http://dip.doe-mbi.ucla.edu`; the somatic nervous system of *Nematode C. elegans* was obtained from `http://ims.dse.ibaraki.ac.jp/research/`.

[17] P. Erdös and A. Rényi, Pulb. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960); B. Bollobás, Random

Graphs (Academic Press, London, 1985).

[18] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, and B.I. Shraiman, Phys. Rev. A **33**, 1141 (1986).

[19] H.E. Stanley and P. Meakin, Nature (London) **335**, 405 (1988).

[20] G. Paladin and A. Vulpiani, Physics Reports **156**, 147 (1987).

[21] R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, J. Phys. A: Math. Gen. **19**, 823 (1986).

[22] B.B. Mandelbrot, J. Fluid Mech. **62**, 331 (1974).

[23] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983); H. Hentschel and I. Procaccia, Physica (Amsterdam) **8D**, 435 (1983); K. Pawelzik and H. G. Schuster, Phys. Rev. A **35**, R481 (1987).

[24] J. White, E. Southgate, J. Thomson, and S. Brenner, Phil. Trans. R. Soc. London B **314**, 1 (1986).

[25] P. Meakin, A. Coniglio, H.E. Stanley, and T.A. Witten, Phys Rev. A **34**, 3325 (1986).

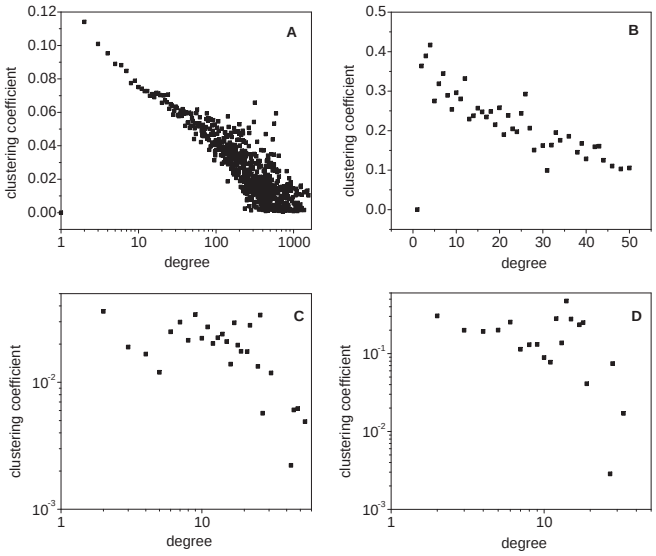

FIG. 1: Plots of the clustering coefficient averaged over all nodes of the same degree versus the degree for selected complex networks. (A) the film actor network, (B) the neural network of *Caenorhabditis elegans*, (C) the protein interaction network of *Helicobacter pylori*, (D) the protein interaction network of *Homo sapiens*. Note that the abscissa of (A), and the abscissa as well as the ordinate of (C) and (D) are in log scale.

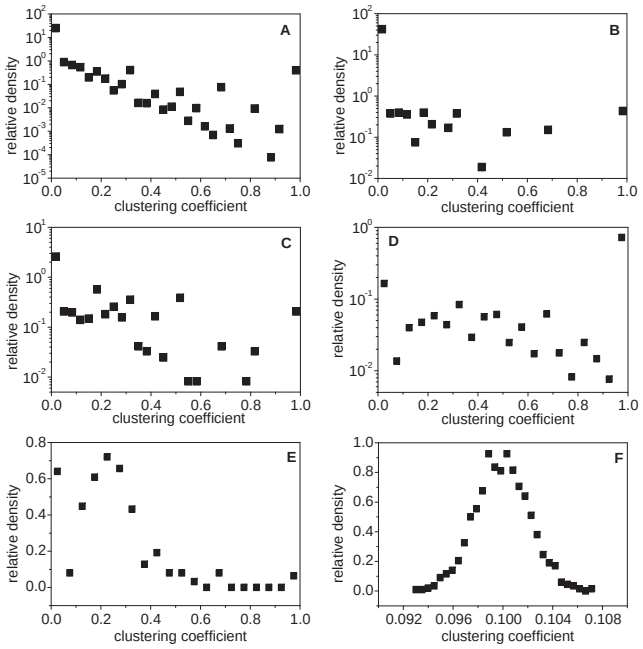

FIG. 2: Probability distributions of the clustering coefficient for selected complex networks as examples. (A) the film actor network, (B) the protein interaction network for *Caenorhabditis elegans*, (C) the metabolic network (in-degree) of *Escherichia coli*, (D) the scientific collaboration network, (E) the neural network, and (F) the random network of 2000 nodes with the connection probability $p = 0.1$. Note that ordinates of (A)-(D) are in log scale for the display purpose.

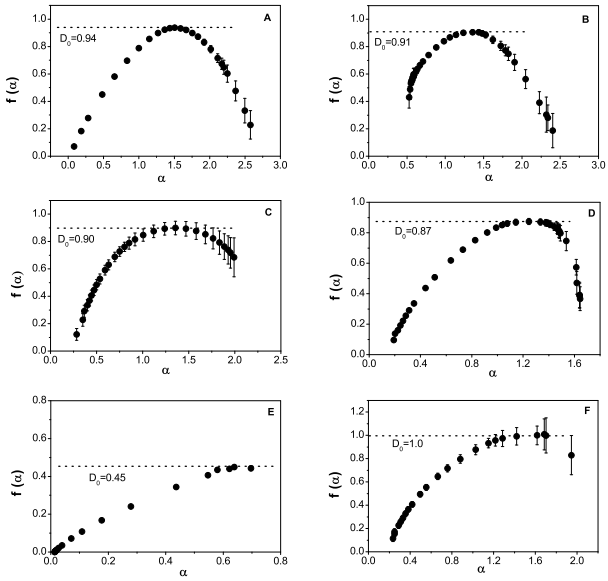

FIG. 3: $f(\alpha)$ versus α for selected complex networks as examples. Error-bars are the root mean square in estimating $f(\alpha)$, and D_0 (the maximum of $f(\alpha)$) is the dimension of the support of the measure. (A) WWW (in-degree), (B) the scientific collaboration network (cond-mat), (C) the metabolic network (out-degree) of *Escherichia coli*, (D) the protein interaction network of *Saccharomyces cerevisiae*, (E) the model of the scale-free network with 10000 nodes, (F) the small-world model of the rewiring probability $p = 0.01$.

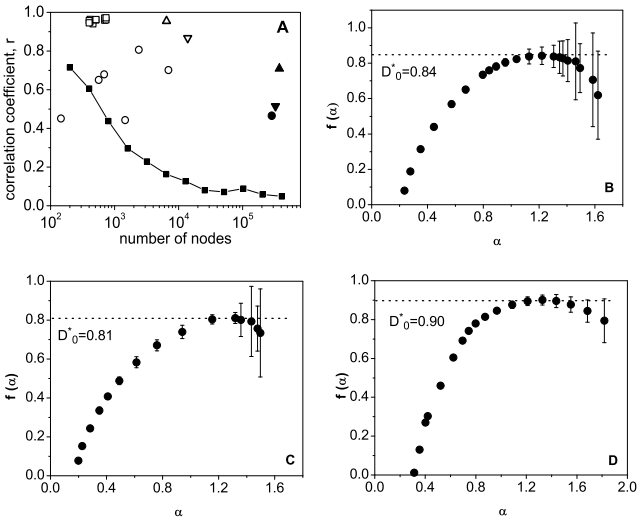

FIG. 4: (A): The linear correlation coefficients r of complex networks plotted against the number of nodes are compared with the random networks of different numbers of nodes: the metabolic networks (□), the Internet of AS level (△), the scientific collaboration network (▽), protein interaction networks (◯), the actor network (▲), WWW (▼), the Internet of router level (•), and the random network (■). For random networks, each node has on average 10 links. (B)-(D): $f(\alpha)$ versus α for selected organisms in the metabolic networks and the Internet of the AS level. D_0^* represents the dimension of the support for the extended clustering coefficient. (B) the metabolic network of *Archaeoglobus fulgidus* (in-degree), (C) the metabolic network of *Caenorhabditis elegans* (in-degree), (D) the Internet of AS level. Note that D_0^* for (B)-(D) are larger than corresponding D_0 for the case of the nearest neighbor, where, from 3B to 3D, $D_0 = 0.79, 0.73, 0.87$, respectively.